BIG DATA – An Overview

Prof. Mahendra S. Patil
Head of Department
Computer Engineering
Atharva College of Engineering
University of Mumbai
Mumbai, India

Jinesh K. Kamdar
Atharva College of Engineering
University of Mumbai
Mumbai, India

Chintan B. Khatri
Atharva College of Engineering
University of Mumbai
Mumbai, India

***ABSTRACT* - Large amounts of data are generated everyday by a variety of sources including cell phone users [1] [5] to military services [2] [13]. This data is so enormous in amount that regular data extraction or querying techniques are useless. The gigantic data is termed as BIG DATA. A very vague term in itself, no proper explanation or measurement is suitable as the limits to big data are very dynamic. For the upper limits to data today might be the lower ones 5-10 years from now. It is very important to clearly understand and decode this term to the fullest as it is of extreme significance in the years to come. This paper aims to explain with subtlety what big data is and why we should know about it. We also go further to provide an overview on the architectures employed, and challenges being faced presently.**

INTRODUCTION

The amount of data generated in the past decade or two collectively is being generated today in a time span of a day or two [4] [5]. The sources responsible for this surge of data are the ever online users of the web continually updating their statuses or uploading some picture or video of a vacation well spent or someone getting married in a lavish ceremony, etc. [1] Not only are these users responsible for the large amounts of data, structured or unstructured, but even companies that are omnisciently collecting data from users by the means of user clicks or camera captures or fingerprints. These sources are adding to an already large collection of data that is laying stagnant most of it useless.

The data used by retail stores a decade ago was hardly relevant to a day's sales. But now with all the data available about a user and his/her history about purchasing a certain product, it

can be used to predict sales and further market analysis. The online retail shopping giants rely on a user's clicks and info to determine what offers to put forward so that the sale is a success [2] [3]. All this is possible because of the unlimited data available freely to be accessed by everyone.

In this paper we attempt to explain the basics of big data. The first section concentrates on the definition of big data, what is it, how is it defined, why is it important. Then we move on to a literature survey i.e. a survey of the work and papers belonging to the field from a specific time period and what the papers concentrated on.

After that we try and explain the architecture to manage and process a typical big data system. We also mention the research objective of this paper and lastly but importantly we list the applications of this relatively new topic of big data.

The real challenge is to manage this data properly as if not done so this imposing data will only slow down the existing technologies in the present day systems by occupying unnecessary space and clogging the system space and memory. The present techniques to manage and process data are not sufficient enough handle such vast amounts of data.

WHAT IS BIG DATA?

"There was five exabytes of information created between the dawns of civilisation through 2003, but that much information is now created every two days, and the pace is increasing."

- Eric Schmidt, former CEO of Google, 2010 [5].

Different organisations define this mysterious term of big data in many different ways. Each trying to modify and adjust the definition according to their specification.

For example, some define big data as a collection of data sets that are typically larger than the processing capability of the existing database technologies [6]. However this definition is subjective and prone to change when viewed by some altogether different person.

Another definition of big data could be the use of technologies to process, analyse, capture and visualise potentially large data sets in a set time frame [8] [12]. With many definitions floating around, it is of utmost importance to properly define this term so that it's understandable to even a lay man.

Hence, big data is nothing but "a collection of a very large amount of data (possibly in terabytes or even bigger) being

generated by numerous users all around the world via different instruments and technologies (such as the web) that is difficult to be handled, processed, analysed and captured/visualised by the existing technologies that process the existing small amounts of data comparatively (even this data is pretty large, but smaller than the term we are focusing on i.e. big data)."

CHARACTERISTICS OF BIG DATA

As explained in the previous part, big data is a very vague term. One needs to really pay attention to understand it. Often while defining a certain term, we tend to overlook other factors or characteristics associated with it that are extremely important for the understanding of the term in question.

Similarly, big data too has certain characteristics that are overlooked while it is being defined. These are the 5 Vs of big data namely, volume, velocity, variety, value and veracity [3] [7] [12] [13]. Let us understand what these terms mean in the context of big data.

Volume: The huge amounts of data being used by companies to improve decision making processes and the data being uploaded and shared by web users.

Variety: By this we mean the large variation in data being generated everywhere, including structured and unstructured data.

Velocity: The speed with which the data is generated is extremely fast. So fast that the present technologies can't handle this speed. Even the refresh speed of data is so high that if not managed properly it could lead to chaos.

Veracity: All the data being generated can't be trusted fully as it may be generated by a malware or some virus somewhere. Veracity is about choosing which data to be trusted from a plethora of it available.

Value: Again the question here arises of the value of data generated. Not all data generated is useful in every scenario. Consider for example the data of a person's loans is not at all useful to an eatery or grocery store.

SYSTEM ARCHITECTURE

Big Data systems usually employ a Lambda Architecture. This architecture specifies a data store that is immutable. An immutable data store removes the update and delete aspects and only allows creation and reading of records. At first this seems to be extremely undesirable as the inability to update data is very frustrating. But after a closer look, we can see that we can delete and update data only in a different way i.e., all the data is recorded with a time stamp so that we know which one is recent.

Consider the following example.
In a mutable data store, say, if a customer prefers USPS, and later the customer prefers, say, DHL. Then this is the case of updating the store as you need to update the database to make it contain the latest record.

However in an immutable data store, say the same customer preferred USPS earlier. Now he prefers DHL. Then in this data store, that customer's preference will be saved based on time. This means both the preferences will be assigned time stamps. The USPS preference with earlier time stamp than the DHL time stamp removing the need of updating and accomplishing the task with just the creation of new data. The lambda architecture supports this [10].

The lambda architecture follows the following setup. It has 3 basic layers [9] [11].

1. Batch layer: Hadoop, an open source framework, stores massive amounts of data. The master data set is stored by the batch layer and arbitrary views are computed using the Hadoop framework. The batch layer doesn't update the data sets frequently resulting in latency. The views are generated from the entire data set.

2. Serving layer: Pre computed views are indexed and exposed to be queried with low latency. The views are queried immediately using open source real time Hadoop query implementations like Cloudera Impala.

3. Speed layer: Batch layer high latency is compensated by the speed layer by computing real time views in distributed stream processing open source solutions like storm and s4. These provide

- Stream processing
- Distributed continuous computation
- Fault tolerance
- Modular design

The need for real time data processing and human fault tolerance drive the decision to implement the lambda architecture. Here are significant benefits from pre computation and recomputation as well as from immutability and human fault tolerance.

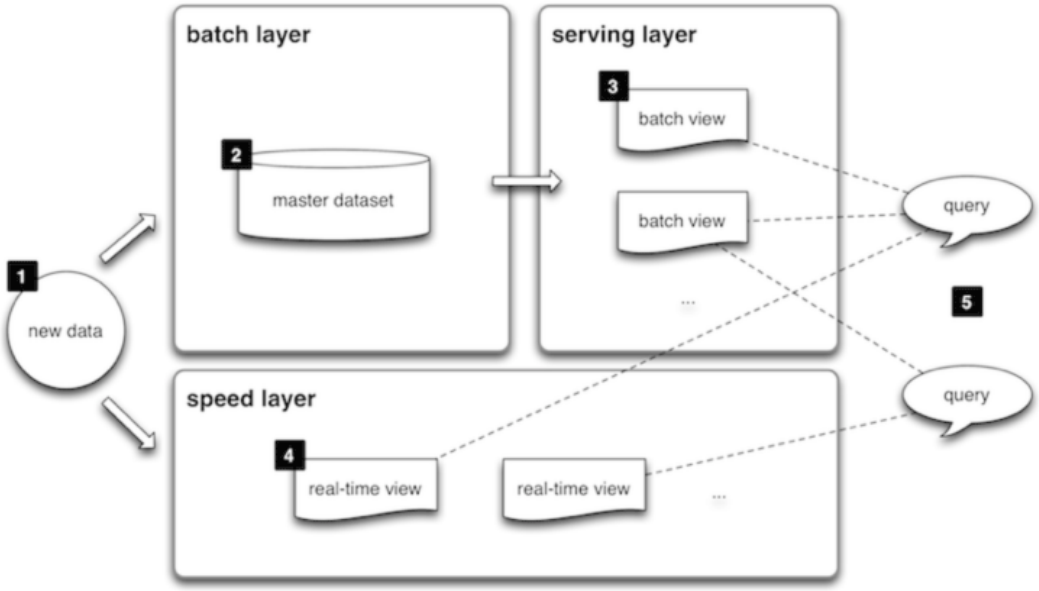

Fig.1 Lambda Architecture for Big Data Systems [10]

LITERATURE SURVEY

Sr. No.	Year	Paper	Description
1	2008	Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society	Growing importance of different computing technologies that can handle big-data.
2	2009	The Pathologies of Big Data	Analyzing how to define big data, an overview of the issues that can arise while handling big data.
3	2011	Big Data: What It Is and Why You Should Care (White Paper)	Big data is a growing challenge. It is the future. Understanding the use of big data.
4	2011-2012	Challenges and Opportunities with Big Data (White Paper)	The humungous amount of data that we are generating has the potential to redefine the way we live. However there are many challenges to be overcome to achieve it.
5	June 2011	Big data: The next frontier for innovation, competition, and productivity	A study of the potential impact of big data in different technological organizations and the economy.
6	October 2011	Are you ready for the era of 'big data'?	Big data has almost arrived. Are we equipped to handle it? How could big data change our decision making process?
7	December 2012	Big Data A New World of Opportunities (NESSI White Paper)	A thorough study on technical, legal, social and marketing aspects of Big Data that have a direct impact on applications, services and software technologies practices. The challenges rising from the use of Big Data and how to overcome them.
8	2012	Entertainment in the Age of Big Data	A fresh perspective about big data. Understanding and exploring relationships as stories.
9	2012	Big Data The Next Big Thing	A look at the opportunity in the services surrounding Big Data with an eye on the future.
10	April 2013	The Data Revolution and Economic Analysis	A study on how new data may impact economic policy and economic research.
11	2013	Survey of Recent Research Progress and Issues in Big Data	An essay on the most recent progress on big data networking and big data.
12	2013	Big Data For Defense And Security	This paper seeks to highlight to defence and security policy-makers the possibilities offered by Big Data, warn about some associated risks and implications of using it.
13	2013	Big Data Analytics	A look at big data analytics, the opportunities it gives rise to, and how big data should be expanded to support analytics.
14	2013	2013 Big Data Survey Research Brief	How companies are planning big data strategies. The technologies they're using as the foundation for these strategies. How they're using big data to solve real business problems.
15	2013	Architecture Framework and Components for the Big Data Ecosystem	This paper discusses a nature of Big Data that may originate from different scientific, industry and social activity domains
16	2013	Dealing with big data: The case of Twitter	An account of experience in working with a big data set, a collection of two billion Dutch tweets.
17	2013	Big Data: New Opportunities and New Challenges	Need to understand big data better. It will entirely change the way we live.

RESEARCH OBJECTIVES

Almost all big companies have a gigantic volume of data, capturing trillions of bytes of information about their customers, their transactions and operations. By the clock, millions of devices such as mobile phones, tablets, laptops, and industrial machines that sense and create data are being added to our world. Big Data keeps getting bigger and bigger every second.

With this paper, we aim to present the various research challenges that a Big Data researcher is facing today.

Big Data can reveal hidden behavioural patterns in humans and help gauge their intentions.

Using big data, if we can find out how a person would react in a particular situation, it would be of immense help to government and private agencies in decision making areas. Data from a hospital patient can help find prevention and cures for a number of deadly diseases.

Management of Big Data in itself is a big challenge. Deciding factors like how much data to store, how secure is it, costing, etc are critical in an organisation.
Big Data is a relatively new phenomenon. It is an evolving concept. New grounds are broken with every new research and a few new challenges arise in the domain.

CONCLUSION

Thus we can conclude with this paper that big data is here to stay. And it will play a very important role in the way the next era of computers shapes up.

Big Data is like an ocean, now it is up to us how to utilise it. It very much exists, but if explored properly, it can yield us rich dividends.

REFERENCES

1. Michael, Katina, and Keith W. Miller. "Big data: New opportunities and new challenges [guest editors' introduction]." Computer 46.6 (2013): 22-24.
2. Bryant, Randal, Randy H. Katz, and Edward D. Lazowska. "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society." (2008): 1-15.
3. Agrawal, Divyakant, et al. "Challenges and opportunities with big data." A community white paper developed by leading researches across the United States (2012).
4. Villars, Richard L., Carl W. Olofson, and Matthew Eastwood. "Big data: What it is and why you should care." White Paper, IDC (2011).
5. Einav, Liran, and Jonathan D. Levin. The data revolution and economic analysis. No. w19035. National Bureau of Economic Research, 2013.
6. Li, Bo. "Survey of Recent Research Progress and Issues in Big Data."
7. Demchenko, Yuri, Canh Ngo, and Peter Membrey. "Architecture Framework and Components for the Big Data Ecosystem." (2013).
8. Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." (2011).
9. http://jameskinley.tumblr.com/post/37398560534/the-lambda-architecture-principles-for-architecting
10. http://www.infoq.com/articles/lambda-architecture-scalable-big-data-solutions
11. http://www.datasciencecentral.com/profiles/blogs/lambda-architecture-for-big-data-systems
12. Big Data A New World of Opportunities, NESSI White Paper, December 2012.
13. Neil Couch and Bill Robins. “BIG DATA FOR DEFENCE AND SECURITY”, 2013.